A new molecular simulation software package – Peking University Drug Design System (PKUDDS) for structure-based drug design

Tingjun Hou and Xiaojie Xu

Department of Chemistry and Molecular Engineering, Peking University, Beijing People's Republic of China

We present a comprehensive molecular simulation program package, the Peking University Drug Design System (PKUDDS), which runs on personal computers. PKUDDS has been developed mainly for computer-aided drug design using the methods of two-dimensional quantitative structure–activity relationships, three-dimensional quantitative structure–activity relationships, molecular docking, and database screening. This study presents an overview of its functionality, especially of methods developed in our group. PKUDDS uses genetic algorithms in molecular docking, conformational analysis, and quantitative structure–lactivity relationships as the most useful optimization technique. A user-friendly graphical interface provides easy access to many functions of PKUDDS. We report some examples of our considerable research using PKUDDS. © 2001 by Elsevier Science Inc.

Keywords: structure-based drug design, computer-aided drug design (CADD), PKUDDS, genetic algorithm (GA), quantitative structure–activity relationships (QSAR), comparative molecular field analysis (CoMFA)

INTRODUCTION

The development of new drugs is a lengthy and expensive process. The first step is to find potential lead compounds with desired biological activity. Computer-aided drug design (CADD) techniques can help increase the pool of interesting structures that can be evaluated. Recent advances have made CADD methods accessible to nonexperts. The rapid increase in computer speed and memory and the decreased cost of personal computers and workstations have brought significant computational resources within the reach of most researchers. Inexpensive computer graphics programs offer improved methods of organizing and visualizing molecular information. Moreover, the algorithms underlying molecular modeling have seen a steady improvement, leading to increasing accuracy in the calculation of molecular properties.

The fundamental assumption of most CADD procedures is that the key biological event, at the molecular level, is the recognition and noncovalent binding of small molecules to specific sites on target biological macromolecules (receptors). Generally, CADD procedures can be divided into two categories: ligand structure-based methods; and receptor structure-based methods. We have incorporated methods such as molecular docking, quantitative structure–activity relationships, and database screening for both of these CADD categories into our modelling package.

OVERVIEW OF PEKING UNIVERSITY DRUG DESIGN SYSTEM (PKUDDS)

We developed the Peking University Drug Design System (PKUDDS) (Figure 1) to provide a convenient method of accessing methods for drug discovery developed in our group. For ease of maintenance and future extensions the system was developed on personal computers to function with Windows 95, Windows 98, or Windows NT operating systems.

PKUDDS provides a powerful simulation capability and a friendly graphical user interface. The computational code and graphical user interface are written with Visual C++.

Color Plate 1 shows the menu bar of PKUDDS containing several menu items, including File, Edit, View, Build, Select, Compute, Analysis, Docking, QSAR, Database, Tools, Setup, and Help. The functions in Figure 1 correspond to different

Color Plates for this article are on pages 474–475.

Corresponding author: X. Xu, Department of Chemistry and Molecular Engineering, Peking University, Beijing 100871, People's Republic of China.

E-mail address: xiaojxu@chem.pku.edu.cn (X. Xu)

Journal of Molecular Graphics and Modelling 19, 455–465, 2001
© 2001 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00094-2

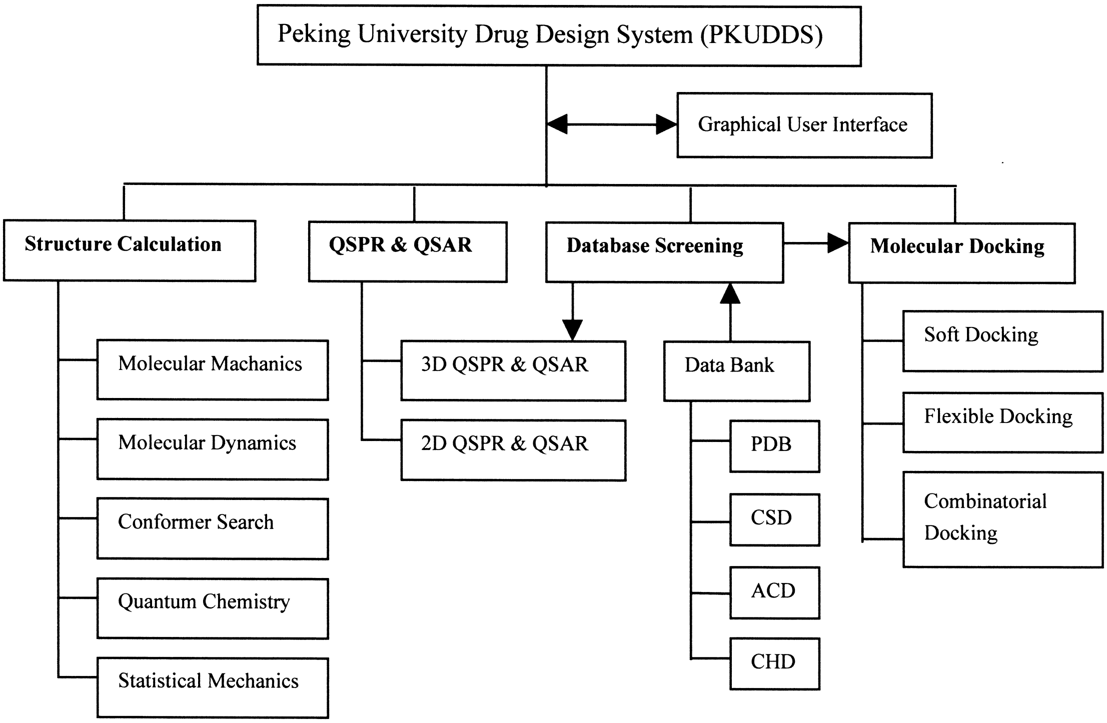


Figure 1. A flow chart for Peking University Drug Design System (PKUDDS).

menu items shown in Color Plate 1. The molecular displays appear in the workspace below the menu bar.

There are two way to generate a molecular model. First, using the tools and editing features under the 'Edit', 'Build,' and 'Select' menu items, we can create a two-dimensional (2D) sketch of a molecule using a mouse, then convert it into a three-dimensional (3D) representation. Second, we can read in atom types and molecular coordinates that have been saved in many file formats including the Sybyl mol2 format, the MDL mol format, the Brookhaven Protein Data Bank format, and the MOPAC dat format. PKUDDS can generate several kinds of rendering styles, including stick, CPK, and ball-and-stick. Color Plate 1 provides an example of a typical molecular representation.

In PKUDDS, the 'Docking' menu item provides access to a soft-docking procedure; the 'QSAR' menu item accesses 2D-QSAR and the enhanced comparative molecular field analysis (CoMFA) based on genetic algorithms (GAs); and the 'Database' menu item allows access to the Chinese herb database (CHD) and related search engines. The package also contains the usual molecular simulation techniques including molecular mechanics, molecular dynamics, conformational search, and quantum chemistry calculations. These are invoked from the sub items in the 'Compute' and 'Analysis' menu items. For molecular mechanics and molecular dynamics simulations, two types of force fields are provided, including MM3[1] and Amber.[2] For conformational searches, GAs are used.[3] Most modules, including molecular mechanics, MOPAC calculations, soft-docking, database searching, and enhanced CoMFA, can be accessed in interactive mode, while other calculations can only be run as background jobs.

PKUDDS METHODOLOGY

Ligand structure-based methods, including quantitative structure–activity relationship (QSAR) methodologies and pharmacophore searches, share the goal of predicting biological activities and devising common pharmacophore models from the physicochemical properties of ligand structures. In most cases, the structure of the receptor is unavailable and the only way to find the SAR models and pharmacophores is from the ligand structures.

Receptor-based methods, including molecular docking and de novo design, seek to find leads by modeling the molecular details of drug action or receptor–ligand interactions. With recent developments in X-ray and NMR techniques, many protein structures have been solved, providing better information about receptor–ligand interactions. With a receptor model in hand, the next step is normally to build or find potential ligands that will fit into the active site model. The key to this step is using 3D information to find or build complementary structures. A crystal structure of the receptor of interest having a ligand bound in the active site offers an ideal place to begin, providing valuable information about the location of important contacts and the conformation of the bound ligand.

The goal of designing PKUDDS was to develop an integrated system on a personal computer that contains all of the functionality necessary for structure-based drug design. Major modules include 2D-QSAR, enhanced CoMFA, molecular docking, and CHD. The source code of the modules in PKUDDS is mainly written in C and C++ languages. For our study, all calculations were performed on a personal computer. Source code and corresponding parameter files in this study, including the 2D-QSAR based on GA, the soft-docking procedure, and the conformational analysis based GA, can be obtained from the author upon request.

2D-QSAR Based on GA

In ligand structure-based methods, traditional two-dimensional quantitative structure–activity relationships (2D-QSAR) analysis is the most widely used and mature technique. QSAR provides a rational basis for understanding mechanisms of biological performance and shows how to improve performance by altering chemical structures of ligands. The underlying assumption is that variations of biological activity of a series of similar structures can be correlated with changes in measured or computed molecular properties of the ligands.

One of the most important and difficult problems in traditional quantitative structure–activity relationship is how to choose the adequate features for building regression models. Recent work has suggested that GAs may be useful reducing the number of features for regression models.[4] Rogers and Hopfinger first applied this method to QSAR analysis,[4] and proved that GAs were very effective tools with advantages over other methods. GA-based QSAR not only can find a group of reliable QSAR models from a large number of samples but is also compatible with nonlinear response surfaces modelled best by higher-order polynomials, splines, and Gaussian models. Moreover, from the analyses of the variables used in the evolution, we may determine which physicochemical properties are most relevant for activity. Consequently, to construct better QSAR models with better predictive abilities, we have added GAs into traditional 2D-QSAR.[5,6]

The basic steps of QSAR based on GAs are as follows:

Creation of the initial population: In genetic algorithms, an individual molecule is represented as a linear string, which plays an analogous role to DNA in evolution. A series of descriptors is randomly added to the string. Every descriptor is expressed using two digits; one digit represents its serial number, and the other represents its function type. The initial population is generated by randomly selecting some number of descriptors from the training set. Then these individuals are ranked according to a fitness score. The "fittest" (elite) individuals (molecules) in the population are retained.

Crossover operation: Once all models in the population have been ranked using the fitness score, the crossover operation is performed repeatedly. In the operation, two good models are probabilistically selected as "parents," with the likelihood of being chosen proportional to a model fitness score; a pair of children is produced by dividing both parents at a randomly chosen point and then joining the pieces together.

Mutation operation: After crossover operation, mutation operations may randomly alter all individuals in the new population and the new model fitness is determined.

Comparison operation: After the crossover and mutation operation, the newly created population and the elite population are compared. If some individuals in the newly created population are better than some individuals in the elite population, these better individuals are copied to the elite population. When the total fitness of the elite population cannot be improved, "convergence" is achieved.

Partial reinitialization: A partial reinitialization procedure is introduced into genetic algorithm by replacing the least fit 50–80% of chromosomes in the population with randomly generated ones after a several steps of crossover and mutation operations. This reduces the likelihood of the GA converging on a local minimum. Generally, 3–6 reinitializations are enough to find all relevant QSAR models.

Upon completion, the models with the highest fitness scores can be obtained. For a population of 200 models, if the data set contains about 20 features, 500–1,000 cycles are usually sufficient to achieve convergence. If the data set has 30 features, 1,000–1,500 operations are usually sufficient. For a typical data set, this process takes 10 min to 1 h for a PC (Pentium 150).

Introduction of GAs to 2D-QSAR is simple, direct, and very effective. Replacing traditional 2D-QSAR with the new procedure based on GA allows the construction of models competitive with, or superior to, standard techniques and makes available additional information that other techniques do not provide.

Enhanced CoMFA Based on GAs

Another important method in QSAR is CoMFA. Since its advent in 1988, CoMFA has been regarded as one of the most powerful tools for three-dimensional quantitative structure–activity relationship analysis (3D-QSAR).[7] The basic assumption in CoMFA is that the observed biological properties can be well understood and correlated with the suitable sampling of the steric and electronic fields surrounding a set of ligands. Because of its wide use, further enhancement of CoMFA will be of considerable benefit to drug researchers.

The major obstacles in generating a CoMFA model on a set of compounds are identifying the bioactive conformation and performing superpositions. Experimental evaluation of the relative energy of the bound conformers of drug molecules supports the argument that it is usually low, but often not the lowest-energy structure that is the biologically relevant one. For relatively rigid compounds, the active conformations will correspond to the lowest-energy conformations. However, for other relatively flexible compounds, selection of appropriate conformations and molecular alignments is more problematic.

For a set of flexible compounds, it is difficult or impossible to manually select appropriate conformations and perform alignments to get the best CoMFA model. For that reason, we introduced GAs into conventional CoMFA to automatically select the conformation for each compound.[8] In the enhanced CoMFA based on GA, the initial population is generated by randomly selecting one conformer for every compound from the training set. Then these individuals are scored using a conventional CoMFA procedure. The cross-validation coefficient (q^2) for every model serves as a fitness score function to evaluate every individual. After crossover, mutation, and comparison operations from the elite population, the models with the highest fitness score can be obtained.

For a population of 50 models, if the data set contains about 20 compounds, 150–300 cycles are usually sufficient to achieve convergence. For a typical data set this process takes 5 to 15 h for a PC (Pentium II 266). The enhanced CoMFA used in this study is under development in our laboratory and is written in C language. In PKUDDS, the enhanced CoMFA based on GA can only be partly run using the graphical interface.

Molecular Docking

Molecular docking can suggest a favorable configuration for two molecules forming a complex system. Molecular docking has been applied to studies of protein–ligand interactions, and structural information from the theoretically modeled complex may help us clarify the mechanism of molecular recognition. Such models can even suggest modifications to lead structures to improve biological activity.

In our group, we have developed different scoring functions for two stages of molecular docking. In the first stage, surface complementarity is considered, while in the second stage only energetics are considered.[9,10]

In the first step, the dot surface is generated using the program written by Connolly.[11] The coordinates of the probe molecule and the target molecule surfaces are randomly rotated and translated. An initial solution is randomly generated containing six variables: three translational degrees of freedom, and three rotational degrees of freedom. The three rotational variables are described by three Euler angles. The position of the target molecule is fixed and the six variables define the orientation of the probe molecule. The initial solution is evaluated using surface complementarity. The evaluation score is composed of two parts: the matching score and penalty score of atomic overlapping.

$$Fitness = Score_{match} - const \times Score_{overlap} \quad (1)$$

where $\mathrm{Score}_{\mathrm{match}}$ is the matching score and $\mathrm{Score}_{\mathrm{overlap}}$ is the penalty score. Const is a coefficient balancing the contributions of the two parts. Const is mainly determined by the dot density, an important parameter of MS program, which is defined as the average number of dots per square angstrom area of both probe and target molecules. In this phase optimization, algorithms, including Monte Carlo simulated annealing, genetic algorithm, and tabu (or taboo) search (TS), are performed repeatedly, and Equation 1 is used to evaluate every solution. After convergence, we arrive at a set of solutions that are then subjected to more detailed searching using each conformation in the first stage.

After the surface complementarity phase, energetics are used as a fitness function. In this stage, only a local search is performed near these binding sites from the surface complementarity. Considering the fast convergence of GA near the best solution, we usually only use GAs in the local search. A set of chromosomes is randomly generated, each one representing an orientation. The fitness score of each chromosome is the interaction energy between the probe and target molecules. Only Van der Waals energy, electrostatic energy, and hydrogen bond energy are considered. The force field used is AMBER.[2] Nonpolar hydrogen atoms are omitted for simplicity and united atom types are introduced to evaluate the interaction energy more efficiently. When the nonbonded interaction energy remains stable in a user-defined region after 20–30 iterations, convergence is achieved. In this stage, different systems are treated in different ways. For protein–protein and some protein–peptide systems, due to the high flexibility of the ligand, it is very difficult to consider conformational flexibility completely. Consequently; for relatively large ligands, only three degrees of translation and three degrees of rotation, and no conformational freedom, are considered. For protein–small-molecule systems, a flexible docking procedure is applied, where the internal conformational flexibility of the ligand is taken into account and some torsional angles are defined as variables in the GA minimizations.

The difference between the two stages is that the first one optimizes the orientations in the whole translational space, but the second stage restrains the translational vectors near the associated sites derived from the first stage.

During the process of molecular recognition between a receptor and its substrate, its potential energy surface is so complicated that it is impossible to determine the associated site by carrying out minimization using gradient methods such as the steepest-descent method and the Gauss-Newton method. Those methods are prone to falling into local potential wells from which they have difficulty escaping. Consequently, some stochastic methods, such as Monte Carlo simulated annealing, have been introduced into studies of molecular association, usually with a more complete potential energetic function. We have used a Simplex method in the minimization procedure and found that it can escape from local minima more easily than Gradient methods. Combined with a random search, Simplex methods can offer a good set of answers to some systems. We have also compared several heuristic algorithms used in previous studies[12] and shown that genetic algorithms and TS were both superior to Monte Carlo simulated annealing algorithm. However, we found that these two algorithms did not perform effectively in all conditions. It is difficult to solve a docking problem completely when using only a single algorithm. With respect to escape from the local minima, TS seems more effective than GA. However, it converges relatively slowly, especially near the best solutions. Consequently, a hybrid algorithm combining GA with TS was proposed.[10] The hybrid algorithm was applied to modelling of protein–peptide and protein–protein complex formation. In our hands our new hybrid algorithm is superior to other heuristic algorithms when used alone. A number of biomolecular systems, including some bound complexes and some unbound complexes, were chosen from the Protein Data Bank (PDB) to test our methods. The results showed that the hybrid minimization algorithm combining GA with TS could successfully find the correct solutions near the observed binding modes for those protein complexes.[10]

Based on the docking procedure developed in our group, a database searching strategy has been proposed. The surface complementarity, energetic considerations, and chemical similarity have been used to rank every molecule in the database. The soft-docking procedure based on surface complementarity can be operated in PKUDDS in interactive mode.

Chinese Herb Database

For database screening based on molecular docking, we used a novel database—the CHD. Unlike other commercial databases, all 3D structures in this database are derived from Chinese herbs. In some respects the CHD offers advantages over other commercial databases in that the structures have come from Chinese medicine and have been proven to be effective for about five thousand years. Additionally, most Chinese herbs have been proven to be very safe or possess only minimal side effects, so the structures in CHD may be safer than those in other databases.

We have built several searchable 3D databases from CHD and applied them in pharmacophore searches and molecular docking. All compounds in CHD are constructed and minimized using molecular mechanics. Moreover, CHD contains additional useful information. including molecular name, CA number, molecular weight, possible medicinal effects, possible biological receptors, toxicity, and relevant references.

All 3D structures in CHD can be displayed and manipulated in PKUDDS. Color Plate 2 = 2 represents one molecule in the subdatabase of HCV agents. In PKUDDS, text matching, molecular weight, and substructure matching searches are available. The CHD database, search engines, and other modules in PKUDDS make up an integrated CHD information system.

APPLICATIONS OF PKUDDS

The structural analysis of ligands and receptors by PKUDDS provides useful information for drug design. Moreover, we have also used PKUDDS to study important processes in the design of new functional materials.[13] We have applied PKUDDS to diverse problems and the results are very encouraging.

Molecular Docking Studies of Two Unbound Complexes

It has been shown that for bound complexes, surface complementarity usually is sufficient to obtain correct binding conformations. However, the ultimate goal of molecular docking is to predict protein–protein and protein–peptide interaction without requiring a complexed crystal structure. Compared with docking of these complexes with crystal structures, calculations for complexes without crystal structures are more difficult. During the formation of a complex, some molecules will undergo conformational changes, so the docking procedure must be sufficiently flexible to manage conformational changes, yet specific enough to identify the correct solution. In some cases, especially when the binding regions between proteins and/or peptides are unknown, complete conformational searches are not tractable. Even using rigid-body approximations, it is very difficult to determine the global minimum using conventional minimization algorithms.

To test our hybrid minimization algorithm and docking procedure, two uncomplexed systems were studied. One study employed an uncomplexed trypsin inhibitor (4PTI in PDB) and an uncomplexed trypsin (3PTN in PDB). The other example used an uncomplexed serine proteinase (2PKA) and an uncomplexed bovine pancreatic trypsin inhibitor (2BPI). The PDB codes of these two cases are 2PTC and 2KAI. All crystallographic water molecules were eliminated from the structures. Some missing hydrogen atoms were added to the complexes using the molecular design software InsightII,[14] with a neutral sp^3 N terminus and a carboxylic (COOH) C terminus assigned at neutral pH. Before calculations, these structures were minimized using the AMBER force field to remove any steric overlap, with the main chain being restrained.

Table 1 summarizes the results for these two cases, with the highest ranked correct prediction illustrated in Figure 2. For 2KAI, we found that the best solution from surface complementarity considerations only did not correspond to the correct docking conformation. After detailed energetic minimization and superimposition with the crystal structure of bound 2KAI, a good solution was found in four of ten solutions from the tabu list. This suggests that, for some unbound complexes, surface complementarity alone cannot be used to reliably dock unbound complexes. Additional energy minimization is needed to filter solutions from the surface complementarity. However, we cannot conclude that the correct solution will have the best energetic complementarity because, in the docking process, we do not adequately consider the flexibility of the systems. For example, in 2KAI, the second solution has the smallest interaction energy, but its root mean square deviation (r.m.s.d.) is larger than 10 Å.

The crystal structures of uncomplexed trypsin (3PTN) and an uncomplexed trypsin inhibitor (4PTI) have been solved separately in different systole forms. A comparison of their structures with the corresponding components of a complex has indicated that relatively large conformational changes have occurred, especially in the trypsin inhibitor. After superimposing only the backbone atoms for 3PTN and 4PTI, the r.m.s.d. to 3PTN is only 0.323 Å; but for 4PTI, the conformational change is relatively large, its r.m.s.d is 1.272 Å. This example

Table 1. The results of molecular docking calculations for two unbound protein complexes

PDB code	Solution number	Rotational eular angles (radian)			Translation vector (Å)			Surface score	Interaction energy (kJ/mol)	RMS (Å)
2PTC	1	30.86	335.75	320.38	14.59	11.36	−12.42	1137.59		2.54
		38.19	9.11	31.66	0.34	−1.15	−1.98		−893.47	2.71
	2	78.03	3.37	325.50	1.16	9.19	16.95	872.78		13.56
		5.16	104.05	312.32	−2.65	1.92	−2.92		−552.01	14.98
	3	128.51	4.13	260.03	−3.62	12.31	16.02	927.55		12.98
		103.25	185.54	105.23	1.50	0.95	0.06		−651.855	11.53
2KAI	1	23.34	9.07	335.17	4.66	−17.26	2.06	973.20		11.98
		34.95	353.03	39.20	−1.01	0.28	1.11		−723.95	10.64
	2	14.97	0.82	316.42	−7.20	7.04	−7.72	726.26		15.98
		134.93	320.42	150.20	−2.00	1.04	−0.85		−965.84	17.83
	3	12.05	1.16	327.56	−1.74	−2.49	−2.48	703.11		14.87
		39.83	334.93	358.23	1.76	0.38	0.33		−232.95	14.99
	4	93.84	1.83	231.47	−10.45	5.13	−1.84	675.05		1.72
		94.83	134.83	338.39	1.06	−0.66	−8.35		−934.87	1.62

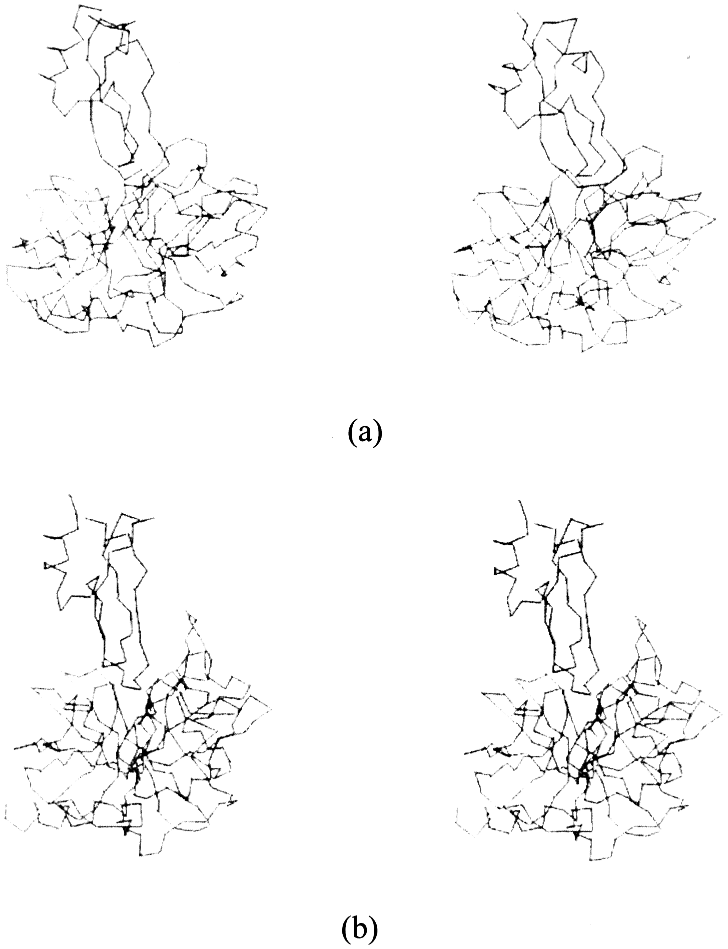
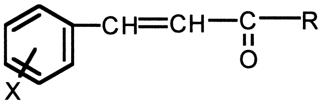
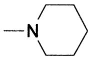


Figure 2. The fitted structure of two unbound complexes. (1). (a) Beta-Trpsin complexes with pancreatic Trypsin inhibitor (2PTC); (b) Kallikrein A complexes with bovine pancreatic trypsin inhibitor. (2) Since the two structures can not be distinguished in the superimposed forms, the fitted structure is moved away from the crystal structure. For each case, the left picture is the crystal structure and the right one is the fitted structure.

is very challenging, and it has been studied extensively by several other docking procedures of Kaichalski-Katzir et al.,[15,16] whose attempt to dock 3PTN and 4PTI was unsuccessful. Our calculation results listed in Table 1 show that the correct binding conformation was found by our hybrid methods. When we superimposed the native complexed crystal structure of 2PTC with our docking result, the r.m.s.d. is 2.54 Å (only backbone atoms were considered in the calculation of r.m.s.d.). From Table 1, it is clear that the best solution from surface complementarity corresponds to the correct solution, but it is no longer significantly better than the rest of the solutions. Consideration of surface complementarity alone is not sufficient to determine which solution is the correct one. After the second stage of energetic minimization, we found that the first solution was more energetically favorable and, in fact, was very close to the correct solution. To understand the influence of the conformational change and test our minimization algorithm further, we docked receptor and ligand from the complexed protein structure together. For the bound and unbound structures, the same parameters were used. However, we found that the calculation results were much different and the best solution from surface complementarity for the bound system was much better than that for the unbound system. When comparing r.m.s.d., the result for the bound system is significantly better than that for the unbound system. The r.m.s.d. for the bound system is only 0.475 Å, much smaller than that for the unbound system. These differences between bound and unbound complexes are mainly derived from the conformational changes during the process of forming the complex. These conformational changes may greatly affect the shape of a molecular surface. Minor changes of molecular surface, especially near the binding site, will greatly affect the docking results. Our methods only implicitly consider the conformation change for these molecules near the binding site, where the conformations do not alter much. In this circumstance, the docking results will produce some deviations from the real complex. However, for 2PTC, the surface did not change significantly in the process of complex formation. We propose that, for most unbound complexes, the best binding conformation may not correspond to that with the best surface complementarity or minimum energy. Instead, some binding conformations with a relatively high surface and energy scores are likely to be found near the correct binding mode. We propose that a kind of consensus score involving both surface

Table 2. Structures of cinnamamides derivatives and experiment and calculated biological activity from equation 3



No.	–R	X	log(1/C) obsd.	log(1/C) calcd[b]	Residue[b]	log(1/C) calcd[c]	Residue[c]
1		3–Cl	0.788	0.510	0.278	0.595	0.193
2		3–F	0.578	0.500	0.078	0.561	0.017
3		4–F	0.458	0.501	−0.043	0.458	0.000
4		4–Br	0.314	0.442	−0.128	0.500	–0.186
5		2,4–Cl	0.664	0.651	0.013	0.623	0.041
6[a]		3,4–Cl	0.550	0.647	−0.097	0.621	−0.071
7		4–Cl	0.606	0.514	0.092	0.596	0.010
8		4–NO_2	0.268	0.324	−0.056	0.314	−0.046
9		3–NO_2	0.324	0.323	0.001	0.310	0.014
10[a]		3–CF_3	0.921	0.815	0.106	0.899	0.022
11		2–CF_3	0.723	0.797	−0.074	0.899	−0.176
12		4–CF_3	0.921	0.819	0.102	0.899	0.022
13		3–OH, 4–OCH_3	−0.272	−0.237	−0.035	−0.272	0.000
14		4–OCH_3	0.218	0.174	0.044	0.270	−0.052
15		3–I	0.320	0.472	−0.152	0.390	−0.070
16		4–OC_2H_5	0.500	0.242	0.258	0.360	0.140
17[a]		4–OC_3H_7–n	0.290	0.332	−0.042	0.348	−0.058
18		4–OC_4H_9–n	0.180	0.400	−0.220	0.268	−0.088
19		3–Cl	0.410	0.586	−0.176	0.651	−0.241
20		3–F	0.495	0.573	−0.078	0.366	0.129
21		4–F	0.495	0.574	−0.079	0.561	−0.066
22		4–Br	0.540	0.517	0.023	0.557	−0.017
23[a]	–HNC_4H_9–i	2,4–Cl_2	0.735	0.732	0.003	0.684	0.051
24		3,4–Cl_2	0.977	0.718	0.259	0.779	0.198
25		4–Cl	0.714	0.583	0.134	0.653	0.061
26		4–CF_3	0.772	0.892	−0.120	0.899	−0.127
27		3–CF_3	0.989	0.890	0.099	0.899	0.090
28[b]		3–Cl	0.620	0.570	0.050	0.600	0.020
29		4–F	0.288	0.562	−0.274	0.366	−0.078
30		4–Br	0.580	0.507	0.073	0.506	0.074
31	–NHC_3H_7–i	2,4–Cl_2	0.600	0.709	−0.109	0.634	−0.034
32		4–Cl	0.801	0.572	0.229	0.603	0.198
33		3,4–Cl_2	0.498	0.704	−0.206	0.629	−0.131
34		4–CF_3	0.899	0.874	0.025	0.899	0.000
35		3–CF_3	0.924	0.875	−0.047	0.899	0.025

[b] These compounds were used as test set and not included in the derivation of equations.
[c] The values of log(1/C) were calculated using Eq. 4.
[d] The values of log(1/C) were calculated using Eq. 18.

complementarity low energy will, in most cases, allow reasonable predictions of the binding mode for an unbound complex.

2D-QSAR Studies of Some Cinnamamides

It is well known that cinnamamide analogs have a wide spectrum of physiological functions, including hypnosis, sedation, anticonvulsant activity, muscle relaxation, local anesthesia, etc. Until now, however, very few studies of the relationship between the chemical structures and biological functions have been reported.

3,4–methylenedioxycinnamoyl piperidide, which possesses distinct anticonvulsant activity, has been identified as a potential anti-epilepsy drug. This compound is a simplified version of poperine II,. Clinical use has shown this compound to have good therapeutic effects in many kinds of epileptic patients, with relatively few side effects. Thirty-five cinnamamide analogs have been synthesized (see Table 2).[17] The chemical structures of these compounds were all modified from 3,4–

Table 3. The parameters used in the QSAR analysis of the data set

Symbol	Explanation
logP	The hydrophobic coefficient of the molecules
π	The hydrophobic coefficient of the substitutes in sites 3, 4, 5
Area	The surface area of the molecules
Vm	The volume of the molecules
Hf	The final heat of formation of the molecules
MW	The molecular weight of the molecules
Density	The density of the molecule
$MR_{2,3,4}$	The total molar refraction in site 2, 3, 4
$\Sigma\sigma$	The hammett σ constant of the substituents on the benzene ring
Fh2o	The aqueous desolvation free energy of the molecules
Apol	Sum of atomic polarizabilities of the molecules
homo, lumo	The energy of home and lumo orbitors of the molecules
Dip, Dip_x, Dip_y, Dip_z	The dipole vector and dipole vector components in x, y, z
Char_N	Atomic net charge of the O atom on the amido group
Char_O	Atomic net charge of the N atom on the amio group

methylendioxycinnamoyl piperidide. These compounds were tested on mice for anticonvulsant activity with the maximal electroshock seizures test (MES) and the ED_{50} was calculated with the Weil method.[17] The potency is defined as $\log(1/ED_{50})$ and is used as the dependent variable in the QSAR study (Table 1).

Molecular geometries of all compounds in Table 2 were modeled and minimized using the InsightII molecular simulation software package.[14] These structures were fully optimized and some quantum-chemical parameters were calculated using the semiempirical AM1 method, available in MOPAC 7.0. Partition coefficients were measured by using the method proposed by Hansch.[18] The aqueous desolvation free energy was calculated from the hydration shell model developed by Hopfinger[19] and the molar refraction came from Zhang et al.[20]

The data set contains 35 compounds and 19 molecular descriptors. The abbreviations for these descriptors are given in Table 3. In our models, five-term and six-term multiple linear regression models were constructed. The use of more than six independent variables for this data set were not considered because of the increasing risk of chance correlations. For this data set, populations with 200 individuals were used and the number of elite populations is defined as 100. After calculations, the 100 best models for five features and four features were obtained. The top 16 models selected from the two elite populations are listed in Table 4. The quality of the models was indicated by SD, F, Q^2, and S_{PRESS}, where n is the number of compounds used in the fit, SD is the standard error of mean, F represents the overall F-statistics for the addition of each successive term, and values in parentheses are the 95% confidence limit of each coefficient.

From the correlation coefficients of descriptors in the top 16 models (Equations 1 – 16) and leave-one-out cross-validations of the models in Table 4, we found Equation 5 to be the best QSAR model with the best predictive ability. The predicted $\log(1/ED_{50})$ values for these 35 compounds are listed in Table 1.

From the statistical analyses of the descriptors used during the evolution process, the principal factors affecting the anticonvulsant activity were determined. The important descriptors were: partition coefficient, molar refractivity, the Hammet σ constant of the substituents on the benzene ring, and the heat of formation of the molecules. To dissect those important factors more closely, these significant parameters, $MR_{2,3,4}$, π, H_f and $\Sigma\sigma$, were used for constructing linear spline models. Correlation studies have shown that these four features were not significantly correlated with each other, indicating that they are all independent features with independent contributions to anticonvulsant activity. The splines used here are truncated power splines and are denoted with angle brackets. For example, $\langle f(x)\text{-}a\rangle$ is equal to zero if the value of (f(x)-a) is negative, otherwise it is equal to (f(x)-a). The regression with splines allows the incorporation of features that do not have a linear effect over their entire range. The terms used in the models were of two functional types: linear polynomials and linear splines. The five-term models were constructed and evaluated in terms of their regression coefficient. QSAR analysis began with a population of 200 random models. The population converged after 850 crossover operations. The best model from the elite population is:

$$\log(1/C) = 0.899 - 0.823(0.70 - MR_{2,3,4}) - 0.008 (H_f + 40.318) - 1.147(0.23 - \Sigma\sigma) - 1.792 (- 0.28 - \pi) \cdot (n = 30 \; r = 0.906 \; F = 28.800 \; Q^2 = 0.744 \; S_{PRESS} = 0.154) \quad (17)$$

The statistics of Equation 17 show that this spline model appears to be much better than the linear regression models in Table 4. The predictivity of this model also appears good.

From this model, the significance and optimum range of each parameter can be obtained. H_f will produce a negative contribution to the anticonvulsant activity when its value is lower than −40.318. High $\Sigma\sigma$ is preferred provided it is no higher than 0.23. We find only eight compounds in which

Table 4. The top 16 QSAR models generated from training set

1. $\log(1/C) = 1.212 + 0.343*\log P - 0.003*Hf - 0.007*Vm + 0.028*Fh2o - 0.014*Dip\_z$
$(n = 30\ \text{Fitness} = 0.862\ SD = 0.621\ F = 13.870\ Q^2 = 0.588\ S_{PRESS} = 0.198)$

2. $\log(1/C) = -0.465 + 0.366*\log P - 0.157*MR_{2,3,4} - 0.002*Hf - 0.004*Vm + 0.023*Fh2o$
$(n = 30\ \text{Fitness} = 0.862\ SD = 0.754\ F = 13.857\ Q^2 = 0.599\ S_{PRESS} = 0.196)$

3. $\text{Log}(1/C) = 0.212 + 0.319*\log P - 0.003*Area - 0.011*Vm - 0.003*Hf + 0.027*Fh2o$
$(n = 30\ \text{Fitness} = 0.857\ SD = 0.813\ F = 13.264\ Q^2 = 0.499\ S_{PRESS} = 0.219)$

4. $\log(1/C) = 1.125 + 0.327*\log P - 0.003*Hf - 0.007*Vm + 0.027*Fh2o$
$(n = 30\ \text{Fitness} = 0.856\ SD = 0.821\ F = 17.095\ Q^2 = 0.600\ S_{PRESS} = 0.191)$

5. $\log(1/C) = 0.566 + 0.403*\log P - 0.330*MR_{2,3,4} - 0.088*lumo - 0.001*Hf + 0.011*Fh2o$
$(n = 30\ \text{Fitness} = 0.855\ SD = 0.624\ F = 13.024\ Q^2 = 0.501\ S_{PRESS} = 0.214)$

6. $\log(1/C) = 0.464 + 0.401*\log P - 0.327*MR_{2,3,4} - 0.001*Hf + 0.017*Fh2o$
$(n = 30\ \text{Fitness} = 0.853\ SD = 0.689\ F = 16.614\ Q^2 = 0.598\ S_{PRESS} = 0.192)$

7. $\log(1/C) = 0.447 + 0.384*MR_{2,3,4} + 0.043*pi + 0.032*Dip - 0.001*Hf$
$(n = 30\ \text{Fitness} = 0.852\ SD = 0.913\ F = 16.580\ Q^2 = 0.569\ S_{PRESS} = 0.197)$

8. $\log(1/C) = -0.302 - 0.600*MR2,3,4 + 0.479*\pi + 0.003*Area - 0.016*Dip\_Z + 0.129*Density$
$(n = 30\ \text{Fitness} = 0.852\ SD = 0.724\ F = 12.709\ Q^2 = 0.471\ S_{PRESS} = 0.225)$

9. $\log(1/C) = -0.449 + 0.012*\Sigma\sigma -0.381*MR_{2,3,4} + 0.429*\pi + 0.030*Dip - 0.001*Hf$
$(n = 30\ \text{Fitness} = 0.852\ SD = 0.834\ F = 12.735\ Q^2 = 0.490\ S_{PRESS} = 0.221)$

10. $\log(1/C) = -0.088 - 0.587*MR_{2,3,4} + 0.487*\pi + 0.003*Area - 0.016*Dip\_z$
$(n = 30\ \text{Fitness} = 0.851\ SD = 0.621\ F = 16.462\ Q^2 = 0.543\ S_{PRESS} = 0.205)$

11. $\log(1/C) = -1.693 - 0.539*MR_{2,3,4} - 0.153*homo + 0.003*Area + 0.002*Dip$
$(n = 30\ \text{Fitness} = 0.851\ SD = 0.754\ F = 12.571\ Q^2 = 0.484\ S_{PRESS} = 0.217)$

12. $\log(1/C) = -1.858 - 0.539*MR_{2,3,4} + 0.443*\pi - 0.170*homo + 0.003*Area$
$(n = 30\ \text{Fitness} = 0.850\ SD = 0.921\ F = 16.349\ Q^2 = 0.531\ S_{PRESS} = 0.207)$

13. $\log(1/C) = -0.294 - 0.587*MR2,3,4 + 0.488*\pi + 0.024*Dip + 0.003*Area$
$(n = 30\ \text{Fitness} = 0.850\ SD = 0.723\ F = 16.28\ Q^2 = 0.529\ S_{PRESS} = 0.208)$

14. $\log(1/C) = -0.688 - 0.341*MR_{2,3,4} + 0.402*\log P - 0.190*lumo - 0.001*Hf$
$(n = 30\ \text{Fitness} = 0.850\ SD = 0.763\ F = 16.259\ Q^2 = 0.595\ S_{PRESS} = 0.193)$

15. $\log(1/C) = -3.212 - 0.253*MR_{2,3,4} - 0.001*Hf + 0.320*\log P - 0.305*homo$
$(n = 30\ \text{Fitness} = 0.850\ SD = 0.723\ F = 16.250\ Q^2 = 0.596\ S_{PRESS} = 0.193)$

16. $\log(1/C) = 0.520 + 0.149*\Sigma\sigma - 0.354*MR_{2,3,4} + 0.397*\pi - 0.001*Hf$
$(n = 30\ \text{Fitness} = 0.850\ SD = 0.604\ F = 16.015\ Q^2 = 0.548\ S_{PRESS} = 0.203)$

electronic effects of substituents on the benzene ring contributed to anticonvulsant activity. A high value of π affords high anticonvulsant activity, but when π is greater than -0.28 there is no longer an increase with the increments of the π values of the substituents on the benzene ring. When the value of π is greater than -0.28, steric and electronic effects become the most important influences on the anticonvulsant activity. High values of $MR_{2,3,4}$ are preferred, as long as the value is below 0.7. Increases in the molar refractivity of the substituents on the benzene ring favor anticonvulsant activity. This conclusion differs from that of the linear regression models, which show that the small substituents on the benzene ring are more favored. This is not inconsistent, because the influent of $MR_{2,3,4}$ is not linear over its entire range. When the inhibitor interacts with its receptor, the steric complementarity is expected to be optimal. We interpret this to mean that the contact area between drug and receptor can increase to a certain point where steric complementarity is optimum, with further increments depressing the activity due to steric hindrance.

In summary, using GA, we generated a group of multiple regression models with high fitness scores. These models were statistically significant and predictive. Steric complementarity and hydrophobic effects are very significant for the biological activity, but the contribution of electronic factors is minimal. We used linear spline models to determine the effective range for the four principal SAR factors.

CoMFA Study of β-Carboline Ligands

The training set used in our study comprised 18 compounds selected directly from literature.[22] All of these compounds were β-carbolines, which possess binding affinity to the benzodiazepine receptor (BzR). Before CoMFA calculations, these compounds were modeled and minimized in InsightII molecular simulation package. The conformational analyses were performed for each compound using the systematic conformational analysis method. Only conformations with energy values within 20 kcal/mol of the global minimum were kept. To simplify calculations, a maximum of 20 conformers were retained for each compound. The partial charges were derived from the CVFF force field.[23]

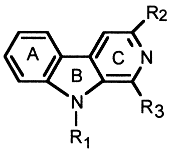
In performing the superimpositions for the CoMFA calculations, eight atoms were used in a least squares fit: six aromatic carbon atoms of the A ring, the indole nitrogen moiety (B ring), and the nitrogen atom (C ring) of β-carbolines and diindoles. The grid used in CoMFA had a resolution of 2.0 Å and the border of the grid is extended about 2Å from the molecules. For this data set, 50 populations were used and the number of elite populations was defined as 10.

Before carrying out calculations using CoMFA based on GA, a conventional CoMFA procedure was first performed to ascertain the appropriate components for every model. In this stage, the lowest-energy conformers for the 18 compounds were selected. A conventional CoMFA calculation was performed, and the cross-validated r^2 (q^2) was calculated for 1–10 components. After a PLS analysis with leave-one-out cross-validation, the CoMFA model with 2 components with q^2 of 0.63 was optimum.

The genetic operator was applied until the total fitness score of the elite populations could not be improved over a period of 15 evolution operations. The convergence criterion was met after 260 operations for 2 components. After convergence, a group of 3D-QSAR models was obtained. Table 5 shows the six best models. It is clear that CoMFA based on GA allows the construction of models superior to standard techniques. The q^2 value of these six models from GA minimization are all higher than those of the model using the lowest-energy conformers, 0.63. Unlike conventional procedures, CoMFA based on GA provides users with multiple models allowing the application of more strict statistical tests to choose the best model.

From the best model in Table 6, it can be seen that not all molecules adopt the lowest-energy conformers. For all 18 molecules, only 8 compounds prefer to adopt the lowest-energy conformers, while other compounds adopt higher energy conformers. The results are not very surprising, because in drug-receptor recognition process, the structures of receptor and ligands will change to gain optimum steric and energetic complementarity.

Table 5. The structures, experimental and predicted biological data for compounds in the *β*-carboline dataset



No.	R^1	R$_2$	R$_3$	Nc^a	PIC$_{50}$ (Actual)	PIC$_{50}$ (Predicted)
1	CO_2CH_3	H	H	2	0.70	1.28
2	$CO_2CH_2CH_3$	H	H	6	0.70	1.42
3	N=C=S	H	H	1	0.90	1.65
4	OCH_2CH_3	H	H	6	1.38	1.53
5	OCH_3	H	H	6	2.70	2.08
6	$O(CH_2)_3CH_3$	H	H	20	1.99	1.81
7	OCH_3	H	H	2	2.09	2.05
8	$O(CH_2)_2CH_3$	H	H	20	1.04	1.57
9	$CO(CH_2)_2CH_3$	H	H	14	0.45	0.53
10	$(CH_2)_3CH_3$	H	H	18	2.39	2.38
11	H	H	H	1	3.21	2.49
12	$CO_2C(CH_3)_3$	H	H	18	1.00	1.37
13	Cl	H	H	1	1.65	1.90
14	NO_2	H	H	1	2.10	1.45
15	$CO_2CH_2C(CH_3)_3$	H	H	8	2.88	2.24
16	CO_2CH_3	H	CH_2CH_3	4	3.88	4.52
17	H	H	CH_2CH_3	4	5.40	5.73
18	H	H	CH_2CH_3	2	4.09	3.44

[a] Nc represents the total conformers for every compounds, in the calculations, the conformers are ranked by their total energy.

Table 6. The predicted biological data from the top six models from the CoMFA based on GA

No.	Conf_1^a	Conf_2^a	Conf_3^a	Conf_4^a	Conf_5^a	Conf_^{6}
1	1	1	2	2	2	2
2	3	3	3	2	2	2
3	1	1	1	1	1	1
4	4	5	4	2	1	1
5	3	1	1	4	4	4
6	15	16	14	17	14	18
7	2	2	2	2	1	1
8	12	12	12	15	12	15
9	11	12	12	9	9	9
10	14	18	18	12	9	9
11	1	1	1	1	1	1
12	2	1	2	2	2	2
13	1	1	1	1	1	1
14	1	1	1	1	1	1
15	1	1	1	2	2	2
16	3	3	2	3	1	1
17	2	2	2	2	3	2
18	1	1	1	1	1	1
R^2	0.87	0.86	0.84	0.81	0.84	0.85
q^2	0.73	0.73	0.72	0.70	0.68	0.68
PRESS^b	0.35	0.42	0.44	0.47	0.44	0.45

[a] The conformer number used in the best model.

[b] PRESS is the predicted sum of squares obtained from the leave-one-out cross-validation method.

Design of New HCV Inhibitors

Hepatitis C virus has been identified as the major causative agent for most cases of non-A, non-B hepatitis. It may establish a chronic infection that persists for decades, which usually results in recurrent and progressively worsening liver inflammation and leads to cirrhosis and hepatocellular carcinoma. To determine the binding site and investigate the interactions between the receptor and inhibitors, our soft-docking and flexible-docking calculations were applied to determine the binding site and the complex structure using available inhibitors. We found that the residues Ser42, Asp81, Lys136, and Gly137 of NS3 protease may contribute to the interactions between the receptor and the inhibitors.

With this binding site information, we performed database searching based on molecular docking to find new potential HCV inhibitors using the CHD. The best hits were chosen and tested for their biological activity. Several molecules among these best hits were proven to possess relatively high biological activities. More detailed information will be discussed in further publications.

ACKNOWLEDGMENTS

We are particularly grateful to Prof. Jay W. Ponder at theWashington University School of Medicine for his distribution of the TINKER molecular simulation package. Some programs for molecular mechanics and molecular dynamics in PKUDDS were obtained from TINKER. This project is supported by the National Natural Sciences Foundation of the People's Republic of China (29992590-2 and 29873005).

REFERENCES

1 Allinger, N.L., Yuh, Y.H., and Lii, J.H. Molecular Mechanics. The MM3 force field for hydrocarbon. 1. *J. Am. Chem. Soc.* 1989, **111**, 8551–8565

2 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. Jr., and Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 1984, **106**, 765–784

3 Wang, J.M., Hou, T.J., Chen, L.R., and Xu, X.J. Conformational analysis of peptides using Monte Carlo simulations combined with the genetic algorithm. *Chemometr. Intell. Lab.* 1999, **45**, 347–351

4 Roger, D., and Hopfinger, A.J. Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* 1994, **34**, 854–866

5 Hou, T.J., Wang, J.M., and Xu, X.J. Applications of genetic algorithms on the structure–activity correlation study of a group of non-nucleoside HIV-1 inhibitors. *Chemometr. Intell. Lab.* 1999, **1–2**, 303–310

6 Hou, T.J., Wang, J.M., Liao, N., and Xu, X.J. Applications of genetic algorithms on the structure–activity relationship analysis of some cinnamamides. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 775–781

7 Cramer, R.D., Patterson, D.E., and Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on steroids to carrier proteins. *J. Am. Chem. Soc.* 1988, **110**, 5959–5967

8 Hou, T.J., Wang, J.M., and Xu, X.J. An enhanced comparative molecular field analysis method using genetic algorithm. *Chinese Chem. Lett.* 1999, **10**, 759–762

9 Wang, J.M., Hou, T.J., and Xu, X.J. Automated docking of peptides and proteins by genetic algorithm. *Chemometr. Intell. Lab.* 1999, **45**, 281–286

10 Hou, T.J., Wang, J.M., and Xu, X.J. Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search. *Protein Eng.* 1999, **12**, 639–647

11 Connolly, M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1983, **221**, 709–713

12 Hou, T.J., Wang, J.M., and Xu, X.J. A comparison of three heuristic algorithms for molecular docking. *Chinese Chem. Lett.* 1999, **10**, 615–618

13 Wang, J.M., Zhang, H., Hou, T.J., and Xu, X.J. Theoretical studies on force titration of amino-group-terminated self-assembled monolayers. *Theochem.–J. Mol. Struc.* 1998, **451**, 295–303 14.InsightII, User's Guide, Molecular Simulation Inc., 1997

15 Katchalski-Katzi, E., Shariv, I., Eisenstein, M., Friesen, A.A., Alfalo, C., and Wodak, S.J. *Proc. Natl. Acad. Sci.* 1992, **89**, 2195–2199

16 Gabb, H.A., Jackson, R.M., and Sternberg, M.J.E. Modelling protein docking using shape complementarity, electrostatic and biochemical information. *J. Mol. Biol.* 1997, **272**, 106–120

17 Li, R.L., and Wang, Y.S. Quantitative structure–anticonvulsant activity relationships of cinnamamides. *Acta Pharmacentica Sinica.* 1986, **21**, 580–585

18 Hansch, C., and Leo, A. *Substituent constants for correlation analysis in chemistry and biology*, John Wiley & Sons, New York, 1979, 49–52.

19 Hopfinger, A.J. *Conformational Properties of Macromolecules*, Academic Press, New York, 1977

20 Zhang, X.H., Li, R.L., and Cai, M.S. Chemical structure–physiological activity relationships in cinnamamides and their analogs: 1. The studies of anticonvulsant activity. *Beijing Med. College Trans.* 1980, **12**, 83–91

21 Mechael, S. Synthetic and computer-assisted analyses of the pharmacophore for the benzodiazepine receptor inverse agonist site. *J. Med. Chem.* **1990**, *33*, 2343–2357

22 Dauber-Osguthorpe, P. Roberts, V.A., Osguthorpe, D.J., Wolff, J., Genest, M., and Hager, A.T. Structure and energetics of ligand binding to proteins: E. coli dihydrofolate reductase-trimethoprim, a drug-receptor system. *Proteins Struct. Funct. Genet.* 1988, **4**, 31–47